OPTIMAL STOPPING TIMES FOR ESTIMATING BERNOULLI PARAMETERS WITH APPLICATIONS TO ACTIVE IMAGING

Safa C. Medin, John Murray-Bruce, and Vivek K Goyal

Boston University Electrical and Computer Engineering Department

ABSTRACT

We address the problem of estimating the parameter of a Bernoulli process. This arises in many applications, including photon-efficient active imaging where each illumination period is regarded as a single Bernoulli trial. We introduce a framework within which to minimize the mean-squared error (MSE) subject to an upper bound on the mean number of trials. This optimization has several simple and intuitive properties when the Bernoulli parameter has a beta prior. In addition, by exploiting typical spatial correlation using total variation regularization, we extend the developed framework to a rectangular array of Bernoulli processes representing the pixels in a natural scene. In simulations inspired by realistic active imaging scenarios, we demonstrate a 4.26 dB reduction in MSE due to the adaptive acquisition, as an average over many independent experiments and invariant to a factor of 3.4 variation in trial budget.

***Index Terms*—** adaptive sensing, Bernoulli processes, beta distribution, computational imaging, conjugate prior, low-light imaging, photon counting, total variation regularization

1. INTRODUCTION

Estimating the parameter of a Bernoulli process is a fundamental problem in statistics and signal processing, and it underlies the relative frequency interpretation of probability [1]. From the outcomes of independent and identically distributed (i.i.d.) binary-valued trials (generically, failure (0) or success (1)), we wish to estimate the probability p of success. Among myriad applications, our interest is raster-scanned active imaging in which a scene patch is periodically illuminated with a pulse, and each period either has a photon-detection event (success) or not (failure) [2]. The probability p has a monotonic relationship with the reflectivity of that patch, and an estimate of p becomes the corresponding image pixel. For efficiency in acquisition time or illumination energy, we are motivated to form the image accurately from a small number of illumination pulses, under conditions where p is small.[1]

Conventional systems are not adaptive. With a fixed number of trials n, the number of successes K is a binomial random variable, and the maximum likelihood (ML) estimate of p is K/n, which has mean-squared error (MSE) of $p(1-p)/n$.[2] As we will show, allowing the number of trials to vary while maintaining an upper bound of n on the average number of trials can result in improved performance under various metrics. We limit our attention here to the MSE of p. First we establish a framework for optimal adaptation of the number of trials for a single Bernoulli process. Then we consider a rectangular array of Bernoulli processes representing a scene in an imaging problem, and we evaluate the inclusion of total variation regularization for exploiting correlations among neighbors. In a simulation with parameters realistic for active optical imaging, we demonstrate a reduction in MSE by a factor of 2.67 (4.26 dB) in comparison to the same regularized reconstruction approach applied without adaptation in numbers of trials.

1.1. Related Work

In statistics, forming a parameter estimate from a number of i.i.d. observations that is dependent on the observations themselves is called *sequential estimation* [4]. Early interest in the sequential estimation problem for a Bernoulli process parameter was inspired by the high relative error of deterministically stopping after n trials when p is small. Specifically, the standard error of the ML estimate is $\sqrt{p(1-p)/n}$, which for small p is unfavorable compared to anything proportional to p. This shortcoming manifests, for example, in the difficulty of distinguishing between two small possible values for p when n is not large.

Haldane [5] observed that if one stops after ℓ successes, the (random) number of trials is informative about p, and a simple unbiased estimate with standard error proportional to p can be found provided that $\ell \geq 3$. Tweedie [6] suggested to call this inverse binomial sampling, but the resulting random variable is now commonly known as negative binomial or Pascal distributed. More recent works have focused on non-MSE performance metrics [7,8], estimation of functions of p [9], estimation from imperfect observations [10], and composite hypothesis testing [11].

First-photon imaging [12] introduced the use of a nondeterministic dwell time to active imaging. This method uses the number of illumination pulses until the first photon is detected to reveal information about reflectivity, setting $\ell = 1$ in the concept of Haldane. Spatial correlations are used to regularize the estimation of the full scene reflectivity image, resulting in good performance from only 1 detected photon per pixel, even when half of the detected photons are attributable to uninformative ambient light. Comparing first-photon imaging to photon-efficient methods with deterministic dwell time [2, 13–21] was an initial inspiration for this work. To the best of our knowledge, no previous paper has explained an advantage or disadvantage from variable dwell time.

1.2. Main Contributions and Outline

This work introduces a novel framework for depicting and understanding stopping rules for estimating a Bernoulli process parameter (Section 2). This framework is not limited to a single error criterion or to Bayesian formulations, but here we limit our attention to MSE

This material is based upon work supported in part by the US National Science Foundation under Grant No. 1422034.

[1] In applications using time-correlated single photon counters, it is recommended to keep p below 0.05 to avoid time skew and missed detections due to detector dead time [3].

[2] Motivations for forming some other estimate of p include a prior distribution for p or a minimax criterion.

of p when a prior distribution for p is known. We develop an optimal data-dependent stopping rule in detail for the case of a Beta prior on p in Section 2.2, extending it to arrays of Bernoulli parameters in Section 3. Numerical results inspired by active imaging applications are presented in Section 4 to validate the proposed schemes. We conclude the paper in Section 5.

2. A SINGLE BERNOULLI PROCESS

Let $\{X_t : t = 1, 2, \ldots\}$ be a Bernoulli process with (unknown random) parameter p, and let $n \in \mathbb{R}^+$ be a *trial budget*. A *(randomized) stopping rule* consists of a sequence of *continuation probability* functions

$$q_t : \{0,1\}^t \to [0,1], \qquad t = 0,\, 1\, \ldots, \tag{1}$$

that give the probability of continuing observations after trial t – based on a biased coin flip independent of the Bernoulli process – as a function of $(X_1, X_2, \ldots, X_t)$. The result is a random number of observed trials T.[3] The stopping rule is said to satisfy the trial budget when $\mathbb{E}[T] \le n$.

Our goal is to minimize the MSE in estimation of p through the design of a stopping rule that satisfies the trial budget and an estimator $\widehat{p}\,(X_1, X_2, \ldots, X_T)$. We will first show that the continuation probability functions can be simplified greatly with no loss of optimality. Then, we will provide results on optimizing the stopping rule under a Beta prior on p.

2.1. Framework for Data-Dependent Stopping

Based on (1), a natural representation of a stopping rule is a binary tree representing all sample paths of the Bernoulli process, with a probability of continuation label at each node. This representation has $2^{t+1} - 1$ labels for observation sequences up to length t. However, the tree can be simplified to a trellis without loss of optimality. Conditioned on observing k successes in m trials, all $\binom{m}{k}$ sequences of length m with k successes are equally likely. Thus, regardless of the prior on p, no improvement can come from having unequal continuation probabilities for tree nodes all representing k successes in m trials. Instead, these nodes should be combined, reducing the tree to a trellis. This representation has $\frac{1}{2}(t+1)(t+2)$ labels for observation sequences up to length t. The continuation probability functions are reduced to a set of probabilities $\{q_{k,m} : m = 0,\, 1\, \ldots;\, k = 0,\, 1,\, \ldots,\, m\}$ for continuing after k successes in m trials, as depicted in Fig. 1.

The conventional use of a fixed number of trials n corresponds to continuation probabilities

$$q_{k,m} = \begin{cases} 1, & m < n; \\ 0, & \text{otherwise.} \end{cases}$$

Regardless of the sample path, one observes exactly n trials, and the number of successes K is a $\mathrm{Binomial}(n, p)$ random variable.

The technique analyzed by Haldane [5] and employed in first-photon imaging [12] with $\ell = 1$ can be expressed with continuation probabilities

$$q_{k,m} = \begin{cases} 1, & k < \ell; \\ 0, & \text{otherwise.} \end{cases}$$

[3]The time T does not satisfy the standard definition of a stopping time because randomness independent of $\{X_t\}$ is allowed to influence the decision of whether to continue the observations. This is to allow the trial budget to be expended exactly.

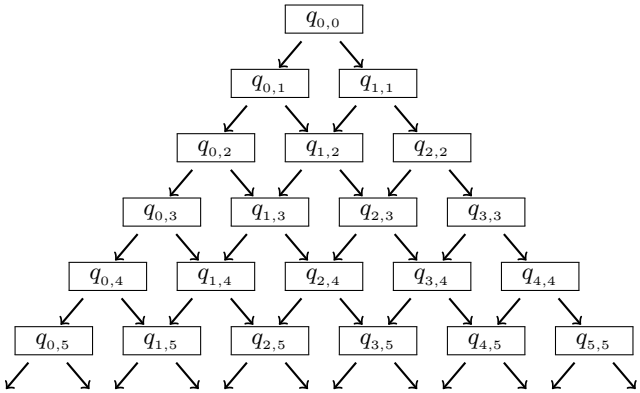


Fig. 1. A trellis showing continuation probabilities for observation sequences up to length 5; $q_{k,m}$ denotes the probability of continuing after observing k successes in m trials.

Observations cease with ℓ successes in M trials, where M is a $\mathrm{NegativeBinomial}(\ell, p)$ random variable.

In general, observations cease with K successes in M trials, where K and M are both random variables. Importantly, the i.i.d. nature of a Bernoulli process makes the pair (K, M) contain all the information that is relevant from the sequence of observations. Similarly to the reduction from tree to trellis, conditioned on $(K, M) = (k, m)$, all sequences of length m with k successes are equally likely, so the specific sequence among these is uninformative about p.

Our method for optimizing the design of continuation probabilities is through analyzing mean Bayes risk reduction from continuation. We define risk function L as *squared error* or *squared loss*

$$L(p, \widehat{p}) = (p - \widehat{p})^2,$$

where p is the Bernoulli parameter and $\widehat{p}$ is the estimate of this parameter. The Bayes risk R is defined as

$$R(p, \widehat{p}) = \mathbb{E}[L(p, \widehat{p})] = \mathbb{E}\left[(p - \widehat{p})^2\right],$$

which in this case is the MSE. Using the minimum MSE (MMSE) estimator, for which $\widehat{p} = \mathbb{E}[P]$, the Bayes risk is the *variance* of the posterior distribution. Thus, key to the optimization is to track posterior variances through the trellis. While this could be done for any prior on p, here we consider only the convenient case of choosing a conjugate prior.

2.2. Analysis Under Beta Prior

The Beta distribution is the conjugate prior for Bernoulli, binomial, and negative binomial distributions. When P has the $\mathrm{Beta}(a, b)$ distribution with probability density function

$$f_P(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1},$$

the posterior after observing k successes in m trials has the $\mathrm{Beta}(a+k, b+m-k)$ distribution. Thus, if we have started with a $\mathrm{Beta}(\alpha, \beta)$ prior, reaching node (k, m) in the trellis implies the posterior distribution is $\mathrm{Beta}(\alpha + k, \beta + m - k)$.

Key facts about $P \sim \mathrm{Beta}(a, b)$ are that its mode is $(a-1)/(a+b-2)$, its mean is $\mathbb{E}[P] = a/(a+b)$, and its variance is

$$\sigma^2_{a,b} = \mathrm{var}(P) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2}$$

Suppose a sequence of trials reaches a node in the trellis corresponding to the posterior distribution $\mathrm{Beta}(a, b)$. The Bayes risk *without* performing an additional trial $R_{\mathrm{stop}}(a, b)$ is the variance $\sigma^2_{a,b}$ given in (2). When one additional trial is performed, the posterior distribution is either $\mathrm{Beta}(a+1, b)$ if the outcome of the additional trial is a success, or $\mathrm{Beta}(a, b+1)$ if the outcome of the additional trial is a failure. Therefore, the mean Bayes risk resulting from continuing with one additional trial is

$$\begin{aligned} R_{\mathrm{cont}}(a, b) &= \mathbb{E}\big[(1-P)\,\sigma^2_{a,b+1} + P\,\sigma^2_{a+1,b}\big] \\ &= \frac{ab}{(a+b)(a+b+1)^2}. \end{aligned} \tag{3}$$

The Bayes risk *reduction* from one additional trial is

$$\begin{aligned} \Delta R(a, b) &= R_{\mathrm{stop}}(a, b) - R_{\mathrm{cont}}(a, b) \\ &= \frac{ab}{(a+b)^2(a+b+1)^2}. \end{aligned} \tag{4}$$

The Bayes risk reduction provides an intuitive – while also theoretically sound – guide to allocating trials. Starting from a $\mathrm{Beta}(\alpha, \beta)$ prior, upon reaching node (k, m), the posterior is $\mathrm{Beta}(\alpha + k, \beta + m - k)$. Then, the Bayes risk reduction from an additional trial would be

$$\Delta R(k, m; \alpha, \beta) = \frac{(\alpha+k)(\beta+m-k)}{(\alpha+\beta+m)^2(\alpha+\beta+m+1)^2}. \tag{5}$$

Let $\Delta_{\mathrm{thresh}} > 0$ denote a specified threshold value for the reduction in Bayes risk that justifies an additional trial. Then the probabilities of continuing at each node of the trellis can be given by

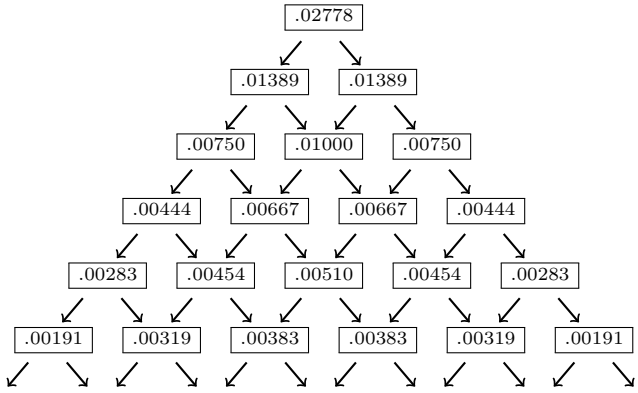
$$q_{k,m} = \begin{cases} 1, & \Delta R(k, m; \alpha, \beta) \geq \Delta_{\mathrm{thresh}}; \\ 0, & \Delta R(k, m; \alpha, \beta) < \Delta_{\mathrm{thresh}}. \end{cases} \tag{6}$$

Figure 2(a) shows an example of a trellis marked with $\Delta R(k, m; \alpha, \beta)$ values, and Figure 2(b) shows the resulting continuation probabilities for $\Delta_{\mathrm{thresh}} = 0.005$.

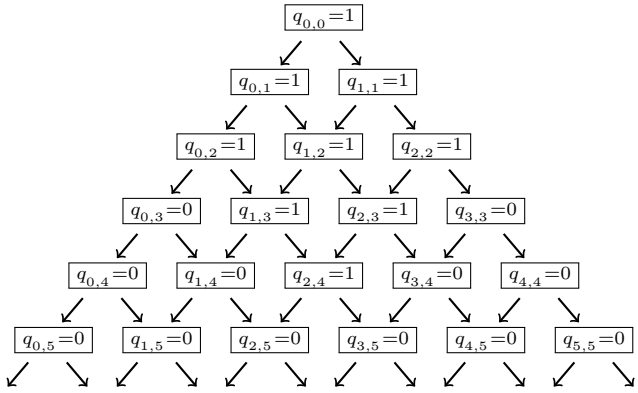
Through a Lagrangian formulation, one can formally establish that sweeping Δ_{thresh} over $[0, \infty)$ is equivalent to varying the trial budget n over $[0, \infty)$, except that we are only achieving the values of $\mathbb{E}[T]$ reachable with $q_{k,m} \in \{0, 1\}$. Intermediate values of $\mathbb{E}[T]$ can be achieved by finding (k^*, m^*) such that $\Delta R(k, m; \alpha, \beta)$ is largest among those below Δ_{thresh} and varying q_{k^*,m^*} over $(0, 1)$.

Notice that for a fixed trellis depth m, the denominator of (5) is fixed, and the numerator of (5) is a product of factors with fixed sum. Thus, from the arithmetic–geometric mean inequality, $\Delta R(k, m; \alpha, \beta)$ is largest where the posterior distribution is most symmetric. This is apparent in the example in Figure 2(b); since we have started with a uniform prior, the center of each row represents a symmetric posterior, and additional observations are most merited near the center of each row. Starting with a highly asymmetric prior ($\alpha \ll \beta$ or $\alpha \gg \beta$), the same principle explains an asymmetry in the optimal continuation probabilities.

The phenomenon of more trials being merited when p is near $\frac{1}{2}$ counteracts the traditional MSE of $p(1-p)/n$ being largest for p near $\frac{1}{2}$. This is illustrated in Figure 3, which shows RMSE as a function of p with and without the optimal stopping. We have optimized for MSE averaged over p and have obtained a modest improvement in this average. A more significant reduction in the worst-case MSE is a by-product of the optimization.



(a) $\Delta R(k, m; \alpha, \beta)$ for $\mathrm{Beta}(1, 1)$ prior.



(b) $q_{k,m}$ values from applying (6) with $\Delta_{\mathrm{thresh}} = 0.005$.

Fig. 2. Trellis representations of the Bayes risk reductions from an additional trial and the resulting continuation probabilities for $\Delta_{\mathrm{thresh}} = 0.005$. A $\mathrm{Beta}(1, 1)$ prior for P (i.e., uniform) is assumed.

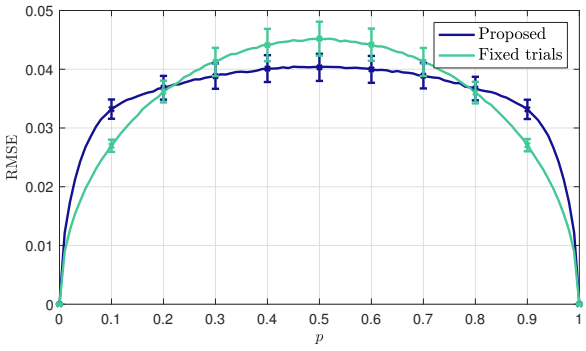
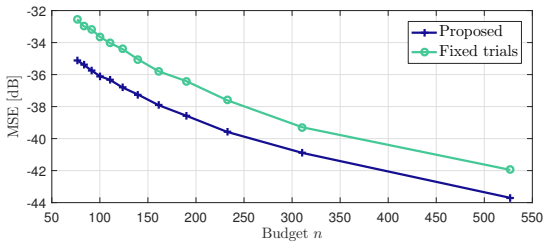
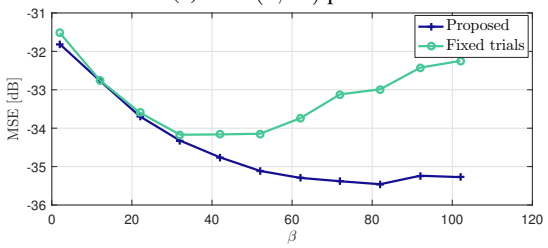


Fig. 3. Dependence of RMSE on the true Bernoulli parameter p, with and without the proposed optimal stopping, when p has a uniform prior and the trial budget is $n = 123$. For p values sampled at multiples of 0.01, the mean from 100 000 experiments is shown; additionally, standard deviations are shown for p values that are multiples of 0.1. The proposed optimal stopping reduces the MSE (averaged over p) from 0.00134 to 0.00129.



(a) $\mathrm{Beta}(2, 80)$ prior.



(b) $n = 83$ and $\mathrm{Beta}(2, \beta)$ prior.

Fig. 4. MSE of pixelwise MMSE estimates for Shepp-Logan Phantom image scaled to $[0.001, 0.201]$.

3. ARRAYS OF BERNOULLI PROCESSES

Active imaging systems typically raster scan the scene by probing patch (i, j), $i = 1, \ldots, N_i$ and $j = 1, \ldots, N_j$, using pulsed illumination. The measured data – used to form an image of the scene – are arrays $[k_{i,j}]_{i,j}$ and $[m_{i,j}]_{i,j}$; i.e., the number of detections (successes) and number of illumination pulses (trials) for each scene patch. Note that the conventional approach of a fixed number of trials makes $m_{i,j} = n$ for all (i, j) and $\{k_{i,j}\}$ random, whereas both $\{k_{i,j}\}$ and $\{m_{i,j}\}$ are random when the proposed approach is applied pixelwise.

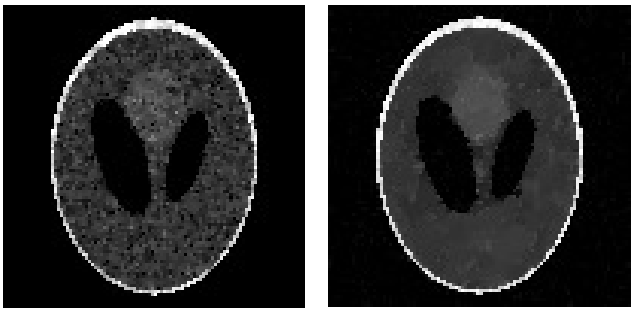
The Bernoulli process generated by probing an individual scene patch (i, j) is typically correlated with the processes of neighboring patches. This can be exploited in the image formation stage through mechanisms inspired by any of various image compression or denoising methods. For this initial demonstration of adaptive acquisition, we apply total variation (TV) regularization. An alternative approach is to attempt to exploit the spatial correlation at the data acquisition stage. This could involve updating the Beta distributions not only for the probed pixel itself, but also for pixels within some neighborhood. The interaction of the adaptive acquisition with regularized estimation is nontrivial and not yet well understood. Therefore, we will defer a study of this approach to future works and focus on the merits of a pixelwise-adaptive data acquisition scheme with TV-regularized reconstructions in the subsequent section.

4. IMAGE ESTIMATION METHODS AND RESULTS

We present simulation results to quantify the performance of the proposed method using MATLAB's built-in Shepp–Logan phantom with 100×100 pixels. The pixel values $p_{i,j}$, rescaled to fall between $(0, 1)$, are used to simulate the underlying Bernoulli process for each pixel to obtain $k_{i,j}$ and $m_{i,j}$. We focus here on comparing conventional fixed number of trials against the data-adaptive stopping rule.

4.1. Stopping rule without TV regularization

We consider pixelwise MMSE estimation using $\mathrm{Beta}(\alpha, \beta)$ priors without regularization: $\widehat{p}_{\mathrm{MMSE}}[i, j] = (k_{i,j} + \alpha)/(m_{i,j} + \alpha + \beta)$.



MSE= 3.4056×10^{-5} MSE= 1.1779×10^{-5}

Fig. 5. Fixed number of trials (left) and optimal stopping (right), showing an MSE improvement of 4.61 dB. The phantom image is rescaled to $[0.001, 0.101]$ and the $\mathrm{Beta}(2, 152)$ prior is assumed. Both images are formed using TV-regularized ML estimation and trial budget $n = 200$.

Table 1. The average reconstruction MSE for the fixed number of trials (Binomial) stopping rule and the adaptive stopping rule, for two different trial budgets n.

Budget	Method	
	Binomial + TV	Adaptive (proposed) + TV
$n = 58$	9.14e-05	**3.43e-05**
$n = 196$	3.37e-05	**1.26e-05**

For one choice of prior, Figure 4(a) shows MSE improvements of at least 2 dB for the data-adaptive stopping rule over the conventional fixed number of trials for the entire range of simulated trial budgets n. For a fixed budget $n = 83$, Figure 4(b) suggests that significant MSE improvements can be gained when the assumed Beta prior becomes more asymmetric (increasing β).

4.2. Stopping rule with TV-regularized image formation

Pixelwise, the ML estimate would be $\widehat{p}_{\mathrm{ML}}[i, j] = k_{i,j}/m_{i,j}$. Reconstruction quality can be improved through the use of TV-regularized ML estimation [2, 22]. In one typical experimental trial shown in Figure 5, the TV-regularized reconstruction from data obtained with the adaptive stopping rule outperforms the conventional method with a 4.61 dB improvement in MSE.

The average MSE improvement is studied by performing 100 independent experiments with TV-regularized image reconstruction when the data is acquired using the Binomial (fixed) and proposed stopping rules. As summarized in Table 1, we observe an overall improvement in reconstruction MSE when adaptive acquisition is used. Interestingly, the improvement factor remains constant at 4.26 dB for both trial budgets.

5. CONCLUSION

We established an optimal stopping framework for estimating Bernoulli parameters that improves the trade-off between MSE and mean number of trials, especially when the prior for the parameter is highly asymmetric. Application of the framework in TV-regularized image estimation was shown to provide significant MSE improvement in active imaging.

6. REFERENCES

[1] T. L. Fine, *Probability and Probabilistic Reasoning for Electrical Engineering*. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

[2] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-efficient computational 3d and reflectivity imaging with single-photon detectors," *IEEE Trans. Comput. Imaging*, vol. 1, pp. 112–125, June 2015.

[3] M. Wahl, "Time-correlated single photon counting (TCSPC)," tech. rep., PicoQuant, Berlin, Germany, 2014.

[4] F. J. Anscombe, "Sequential estimation," *J. Roy. Statist. Soc. Ser. B*, vol. 15, no. 1, pp. 1–29, 1953.

[5] J. B. S. Haldane, "On a method of estimating frequencies," *Biometrika*, vol. 33, pp. 222–225, Nov. 1945.

[6] M. C. K. Tweedie, "Inverse statistical variates," *Nature*, vol. 155, p. 453, Apr. 14, 1945.

[7] P. Cabilio and H. Robbins, "Sequential estimation of p with squared relative error loss," *Proc. Nat. Acad. Sci. USA*, vol. 72, pp. 191–193, Jan. 1975.

[8] P. Cabilio, "Sequential estimation in Bernoulli trials," *Ann. Statist.*, vol. 5, pp. 342–356, Mar. 1977.

[9] S. L. Hubert and R. Pyke, "Sequential estimation of functions of p for Bernoulli trials," in *Game Theory, Optimal Stopping, Probability and Statistics*, vol. 35 of *Lecture Notes-Monograph Series*, pp. 263–294, Institute of Mathematical Statistics, 2000.

[10] P. M. Djurić and Y. Huang, "Estimation of a Bernoulli parameter p from imperfect trials," *IEEE Signal Process. Lett.*, vol. 7, pp. 160–163, June 2000.

[11] D. Ciuonzo, A. De Maio, and P. Salvo Rossi, "A systematic framework for composite hypothesis testing of independent Bernoulli trials," *IEEE Signal Process. Lett.*, vol. 22, pp. 1249–1253, Sept. 2015.

[12] A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F. N. C. Wong, J. H. Shapiro, and V. K. Goyal, "First-photon imaging," *Science*, vol. 343, no. 6166, pp. 58–61, 2014.

[13] N. J. Krichel, A. McCarthy, and G. S. Buller, "Resolving range ambiguity in a photon counting depth imager operating at kilometer distances," *Opt. Express*, vol. 18, no. 9, pp. 9192–9206, 2010.

[14] P. A. Morris, R. S. Aspden, J. E. C. Bell, R. W. Boyd, and M. J. Padgett, "Imaging with a small number of photons," *Nat. Commun.*, vol. 6, Jan. 5, 2015. doi: 10.1038/ncomms6913.

[15] D. Shin, J. H. Shapiro, and V. K. Goyal, "Single-photon depth imaging using a union-of-subspaces model," *IEEE Signal Process. Lett.*, vol. 22, pp. 2254–2258, Dec. 2015.

[16] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Lidar waveform-based analysis of depth images constructed using sparse single-photon data," *IEEE Trans. Image Process.*, vol. 25, pp. 1935–1946, May 2016.

[17] D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappa, V. K. Goyal, F. N. C. Wong, and J. H. Shapiro, "Photon-efficient imaging with a single-photon camera," *Nat. Commun.*, vol. 7, June 24, 2016. doi: 10.1038/ncomms12046.

[18] D. Shin, J. H. Shapiro, and V. K. Goyal, "Performance analysis of low-flux least-squares single-pixel imaging," *IEEE Signal Process. Lett.*, vol. 23, pp. 1756–1760, Dec. 2016.

[19] L. Mertens, M. Sonnleitner, J. Leach, M. Agnew, and M. J. Padgett, "Image reconstruction from photon sparse data," *Sci. Rep.*, vol. 7, Feb. 7, 2017. doi: 10.1038/srep42164.

[20] J. Rapp and V. K. Goyal, "A few photons among many: Unmixing signal and noise for photon-efficient active imaging," *IEEE Trans. Comput. Imaging*, vol. 3, pp. 445–459, Sept. 2017.

[21] Y. Altmann, R. Aspden, M. Padgett, and S. McLaughlin, "A Bayesian approach to denoising of single-photon binary images," *IEEE Trans. Comput. Imaging*, vol. 3, pp. 460–471, Sept. 2017.

[22] C. Louchet and L. Moisan, "Total variation denoising using posterior expectation," in *Proc. 16th European Signal Process. Conf.*, pp. 1–5, Aug. 2008.