Convergence of Standard Gradient Descent 1

Problem Setup

We consider the minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function. The standard gradient descent update rule is:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \tag{1}$$

where $\alpha > 0$ is the step size (learning rate), and $\nabla f(x_k)$ is the gradient of f at x_k with respect to the standard Euclidean inner product $\langle u, v \rangle = u^T v$. We use the standard Euclidean norm $\|u\| = \sqrt{\langle u, u \rangle}$.

Assumptions

1. **Convexity of f**: For any $x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{2}$$

2. **L-smoothness (Lipschitz continuous gradient)**: The gradient ∇f is L-Lipschitz continuous with respect to the Euclidean norm. That is, there exists a constant $L > 0$ such that:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n \tag{3}$$

   This assumption implies the following inequality (Descent Lemma):

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \tag{4}$$

3. **Existence of a minimizer**: There exists $x^* \in \mathbb{R}^n$ such that $f(x^*) = f^* = \min_{x \in \mathbb{R}^n} f(x)$. For convex f, this implies $\nabla f(x^*) = 0$.

Convergence Proof

Step 1: Bounding the function decrease in one step. We use the Descent Lemma (4) with $x = x_k$ and $y = x_{k+1} = x_k - \alpha \nabla f(x_k)$. Then $y - x = -\alpha \nabla f(x_k)$.

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), -\alpha \nabla f(x_k) \rangle + \frac{L}{2}\| - \alpha \nabla f(x_k)\|^2 \\
&= f(x_k) - \alpha \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2 \\
&= f(x_k) - \alpha \|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2 \\
&= f(x_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x_k)\|^2
\end{aligned}$$

To guarantee descent, we choose the step size α such that $1 - \frac{L\alpha}{2} \geq 0$, i.e., $\alpha \leq \frac{2}{L}$. A common choice is $\alpha = \frac{1}{L}$. With this choice:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{L}\left(1 - \frac{L(1/L)}{2}\right)\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{L}\left(1 - \frac{1}{2}\right)\|\nabla f(x_k)\|^2 \\ f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \qquad (5) \end{aligned}$$

This shows that the function value decreases at each step, provided $\nabla f(x_k) \neq 0$.

Step 2: Analyzing the distance to the optimum. Consider the squared Euclidean distance from x_{k+1} to x^*:

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha\nabla f(x_k) - x^*\|^2 \\ &= \|(x_k - x^*) - \alpha\nabla f(x_k)\|^2 \\ &= \|x_k - x^*\|^2 - 2\langle x_k - x^*, \alpha\nabla f(x_k)\rangle + \|\alpha\nabla f(x_k)\|^2 \\ &= \|x_k - x^*\|^2 - 2\alpha\langle\nabla f(x_k), x_k - x^*\rangle + \alpha^2\|\nabla f(x_k)\|^2 \end{aligned}$$

Step 3: Using convexity. From the convexity inequality (2), substitute $y = x^*$:

$$f(x^*) \geq f(x_k) + \langle\nabla f(x_k), x^* - x_k\rangle$$

Rearranging gives:

$$\langle\nabla f(x_k), x_k - x^*\rangle \geq f(x_k) - f(x^*) = f(x_k) - f^* \qquad (6)$$

Step 4: Combining the results. Substitute inequality (6) into the expression for the distance:

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha(f(x_k) - f^*) + \alpha^2\|\nabla f(x_k)\|^2$$

Now, use the step size $\alpha = 1/L$ and the function decrease inequality (5). From (5), we have $\|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f(x_{k+1}))$.

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - \frac{2}{L}(f(x_k) - f^*) + \frac{1}{L^2}\|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - \frac{2}{L}(f(x_k) - f^*) + \frac{1}{L^2}[2L(f(x_k) - f(x_{k+1}))] \\ &= \|x_k - x^*\|^2 - \frac{2}{L}(f(x_k) - f^*) + \frac{2}{L}(f(x_k) - f(x_{k+1})) \\ &= \|x_k - x^*\|^2 - \frac{2}{L}\left[(f(x_k) - f^*) - (f(x_k) - f(x_{k+1}))\right] \\ &= \|x_k - x^*\|^2 - \frac{2}{L}(f(x_{k+1}) - f^*) \end{aligned}$$

Step 5: Telescoping sum. Let $\delta_k = \|x_k - x^*\|^2$ (squared Euclidean distance) and $\varepsilon_k = f(x_k) - f^*$ (function error). The inequality becomes:

$$\delta_{k+1} \leq \delta_k - \frac{2}{L}\varepsilon_{k+1}$$

or

$$\varepsilon_{k+1} \leq \frac{L}{2}(\delta_k - \delta_{k+1})$$

Summing this inequality from $k = 0$ to $K - 1$:

$$\begin{aligned}
\sum_{k=0}^{K-1} \varepsilon_{k+1} &\leq \sum_{k=0}^{K-1} \frac{L}{2}(\delta_k - \delta_{k+1}) \\
\sum_{k=1}^{K} \varepsilon_k &\leq \frac{L}{2}\left(\sum_{k=0}^{K-1}(\delta_k - \delta_{k+1})\right) \\
&= \frac{L}{2}(\delta_0 - \delta_K) \quad \text{(Telescoping sum)}
\end{aligned}$$

Since $\delta_K = \|x_K - x^*\|^2 \geq 0$, we have $\delta_0 - \delta_K \leq \delta_0 = \|x_0 - x^*\|^2$. Thus:

$$\sum_{k=1}^{K} \varepsilon_k \leq \frac{L}{2}\|x_0 - x^*\|^2 \tag{7}$$

Step 6: Obtaining the convergence rate. From Step 1, we know that $f(x_{k+1}) \leq f(x_k)$, so the sequence $\varepsilon_k = f(x_k) - f^*$ is non-increasing ($\varepsilon_{k+1} \leq \varepsilon_k$). Therefore:

$$K \cdot \varepsilon_K = K(f(x_K) - f^*) \leq \sum_{k=1}^{K} \varepsilon_k$$

Combining this with inequality (7):

$$K\varepsilon_K \leq \frac{L}{2}\|x_0 - x^*\|^2$$

From this, we obtain the convergence rate:

$$f(x_K) - f^* = \varepsilon_K \leq \frac{L\|x_0 - x^*\|^2}{2K} \tag{8}$$

Conclusion

For a convex and L-smooth function f, the standard gradient descent method with step size $\alpha = 1/L$ converges in function value to the minimum f^* with a rate of $O(1/K)$. That is, $f(x_K) \to f^*$ as $K \to \infty$.

Derivation of the Preconditioned Gradient $\nabla_P f(x)$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Let P be a symmetric positive definite matrix defining the P-inner product :

$$\langle u, v \rangle_P = u^T P v = \langle Pu, v \rangle$$
$$\|u\|_P = \sqrt{\langle u, u \rangle_P} = \sqrt{u^T P u}$$

The differential of f at x, denoted $df_x(h)$, is a linear functional representing the best linear approximation of the change $f(x+h) - f(x)$ for a small displacement h. By the Riesz Representation Theorem, this linear functional df_x can be represented via an inner product with a unique vector. The specific vector depends on the chosen inner product:

1. Using the **standard inner product** $\langle \cdot, \cdot \rangle$, there exists a unique vector, the **standard gradient** $\nabla f(x)$, such that for all $h \in \mathbb{R}^n$:

$$df_x(h) = \langle \nabla f(x), h \rangle$$

2. Using the P**-inner product** $\langle \cdot, \cdot \rangle_P$, there exists a unique vector, the **preconditioned gradient** $\nabla_P f(x)$, such that for all $h \in \mathbb{R}^n$:

$$df_x(h) = \langle \nabla_P f(x), h \rangle_P$$

Since both expressions represent the same differential $df_x(h)$, they must be equal:

$$\langle \nabla f(x), h \rangle = \langle \nabla_P f(x), h \rangle_P$$

Using the definition $\langle u, v \rangle_P = \langle Pu, v \rangle$:

$$\langle \nabla f(x), h \rangle = \langle P(\nabla_P f(x)), h \rangle$$

Rearranging the terms:

$$\langle \nabla f(x) - P\nabla_P f(x), h \rangle = 0$$

This must hold for all h, which implies the vector inside the inner product must be zero:

$$\nabla f(x) - P\nabla_P f(x) = 0$$

Solving for $\nabla_P f(x)$ (using the invertibility of P):

$$P\nabla_P f(x) = \nabla f(x)$$

$$\boxed{\nabla_P f(x) = P^{-1} \nabla f(x)} \tag{9}$$

This gives the explicit relationship between the preconditioned gradient and the standard gradient.

Convergence of Preconditioned Gradient Descent 1

Problem Setup

We consider the minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function. The preconditioned gradient descent update rule is:

$$x_{k+1} = x_k - \alpha \nabla_P f(x_k) = x_k - \alpha P^{-1} \nabla f(x_k) \tag{10}$$

where $\alpha > 0$ is the step size (learning rate).

Assumptions

1. **Convexity of f**: For any $x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle = f(x) + \langle \nabla_P f(x), y - x \rangle_P \tag{11}$$

2. **L_P-smoothness with respect to the P-norm**: We assume that the preconditioned gradient $\nabla_P f$ is L_P-Lipschitz continuous with respect to the P-norm. That is, there exists a constant $L_P > 0$ such that:

$$\|\nabla_P f(x) - \nabla_P f(y)\|_P \leq L_P \|x - y\|_P \quad \forall x, y \in \mathbb{R}^n \tag{12}$$

   This assumption is equivalent to the following inequality (the P-Descent Lemma):

$$f(y) \leq f(x) + \langle \nabla_P f(x), y - x \rangle_P + \frac{L_P}{2} \|y - x\|_P^2 \tag{13}$$

3. **Existence of a minimizer**: There exists $x^* \in \mathbb{R}^n$ such that $f(x^*) = f^* = \min_{x \in \mathbb{R}^n} f(x)$. For convex f, this implies $\nabla f(x^*) = 0$, and consequently $\nabla_P f(x^*) = P^{-1} 0 = 0$.

Convergence Proof

Step 1: Bounding the function decrease in one step. We use the P-Descent Lemma (13) with $x = x_k$ and $y = x_{k+1} = x_k - \alpha \nabla_P f(x_k)$. Then

$y - x = -\alpha \nabla_P f(x_k)$.

$$
\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \langle \nabla_P f(x_k), -\alpha \nabla_P f(x_k) \rangle_P + \frac{L_P}{2} \| - \alpha \nabla_P f(x_k) \|_P^2 \\
&= f(x_k) - \alpha \langle \nabla_P f(x_k), \nabla_P f(x_k) \rangle_P + \frac{L_P \alpha^2}{2} \| \nabla_P f(x_k) \|_P^2 \\
&= f(x_k) - \alpha \| \nabla_P f(x_k) \|_P^2 + \frac{L_P \alpha^2}{2} \| \nabla_P f(x_k) \|_P^2 \\
&= f(x_k) - \alpha \left(1 - \frac{L_P \alpha}{2} \right) \| \nabla_P f(x_k) \|_P^2
\end{aligned}
$$

To guarantee descent, we choose the step size α such that $1 - \frac{L_P \alpha}{2} \ge 0$, i.e., $\alpha \le \frac{2}{L_P}$. A standard choice is $\alpha = \frac{1}{L_P}$. With this choice:

$$
\begin{aligned}
f(x_{k+1}) &\le f(x_k) - \frac{1}{L_P} \left(1 - \frac{L_P(1/L_P)}{2} \right) \| \nabla_P f(x_k) \|_P^2 \\
&= f(x_k) - \frac{1}{L_P} \left(1 - \frac{1}{2} \right) \| \nabla_P f(x_k) \|_P^2 \\
f(x_{k+1}) &\le f(x_k) - \frac{1}{2L_P} \| \nabla_P f(x_k) \|_P^2 \qquad (14)
\end{aligned}
$$

This shows that the function value decreases at each step, provided $\nabla_P f(x_k) \neq 0$.

Step 2: Analyzing the distance to the optimum in P-norm. Consider the squared P-norm of the distance from x_{k+1} to x^*:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_P^2 &= \|x_k - \alpha \nabla_P f(x_k) - x^*\|_P^2 \\
&= \|(x_k - x^*) - \alpha \nabla_P f(x_k)\|_P{}_P^2 \\
&= \|x_k - x^*\|_P^2 - 2\langle x_k - x^*, \alpha \nabla_P f(x_k) \rangle_P + \|\alpha \nabla_P f(x_k)\|_P^2 \\
&= \|x_k - x^*\|_P^2 - 2\alpha \langle \nabla_P f(x_k), x_k - x^* \rangle_P + \alpha^2 \|\nabla_P f(x_k)\|_P^2
\end{aligned}
$$

Step 3: Using convexity. From the convexity inequality (11), substitute $y = x^*$:

$$
f(x^*) \ge f(x_k) + \langle \nabla_P f(x_k), x^* - x_k \rangle_P
$$

Rearranging gives:

$$
\langle \nabla_P f(x_k), x_k - x^* \rangle_P \ge f(x_k) - f(x^*) = f(x_k) - f^* \qquad (15)
$$

Step 4: Combining the results. Substitute inequality (15) into the expression for the distance:

$$
\|x_{k+1} - x^*\|_P^2 \le \|x_k - x^*\|_P^2 - 2\alpha (f(x_k) - f^*) + \alpha^2 \|\nabla_P f(x_k)\|_P^2
$$

Now, use the step size $\alpha = 1/L_P$ and the function decrease inequality (14). From (14), we have $\|\nabla_P f(x_k)\|_P^2 \leq 2L_P(f(x_k) - f(x_{k+1}))$.

$$\begin{aligned}
\|x_{k+1} - x^*\|_P^2 &\leq \|x_k - x^*\|_P^2 - \frac{2}{L_P}(f(x_k) - f^*) + \frac{1}{L_P^2}\|\nabla_P f(x_k)\|_P^2 \\
&\leq \|x_k - x^*\|_P^2 - \frac{2}{L_P}(f(x_k) - f^*) + \frac{1}{L_P^2}[2L_P(f(x_k) - f(x_{k+1}))] \\
&= \|x_k - x^*\|_P^2 - \frac{2}{L_P}(f(x_k) - f^*) + \frac{2}{L_P}(f(x_k) - f(x_{k+1})) \\
&= \|x_k - x^*\|_P^2 - \frac{2}{L_P}\left[(f(x_k) - f^*) - (f(x_k) - f(x_{k+1}))\right] \\
&= \|x_k - x^*\|_P^2 - \frac{2}{L_P}(f(x_{k+1}) - f^*)
\end{aligned}$$

Step 5: Telescoping sum. Let $\delta_k^P = \|x_k - x^*\|_P^2$ (squared P-distance) and $\varepsilon_k = f(x_k) - f^*$ (function error). The inequality becomes:

$$\delta_{k+1}^P \leq \delta_k^P - \frac{2}{L_P}\varepsilon_{k+1}$$

or

$$\varepsilon_{k+1} \leq \frac{L_P}{2}(\delta_k^P - \delta_{k+1}^P)$$

Summing this inequality from $k = 0$ to $K - 1$:

$$\begin{aligned}
\sum_{k=0}^{K-1} \varepsilon_{k+1} &\leq \sum_{k=0}^{K-1} \frac{L_P}{2}(\delta_k^P - \delta_{k+1}^P) \\
\sum_{k=1}^{K} \varepsilon_k &\leq \frac{L_P}{2}\left(\sum_{k=0}^{K-1}(\delta_k^P - \delta_{k+1}^P)\right) \\
&= \frac{L_P}{2}(\delta_0^P - \delta_K^P) \quad \text{(Telescoping sum)}
\end{aligned}$$

Since $\delta_K^P = \|x_K - x^*\|_P^2 \geq 0$, we have $\delta_0^P - \delta_K^P \leq \delta_0^P = \|x_0 - x^*\|_P^2$. Thus:

$$\sum_{k=1}^{K} \varepsilon_k \leq \frac{L_P}{2}\|x_0 - x^*\|_P^2 \tag{16}$$

Step 6: Obtaining the convergence rate. From Step 1, we know that $f(x_{k+1}) \leq f(x_k)$, so the sequence $\varepsilon_k = f(x_k) - f^*$ is non-increasing ($\varepsilon_{k+1} \leq \varepsilon_k$). Therefore:

$$K \cdot \varepsilon_K = K(f(x_K) - f^*) \leq \sum_{k=1}^{K} \varepsilon_k$$

Combining this with inequality (16):

$$K\varepsilon_K \leq \frac{L_P}{2}\|x_0 - x^*\|_P^2$$

From this, we obtain the convergence rate:

$$f(x_K) - f^* = \varepsilon_K \le \frac{L_P \|x_0 - x^*\|_P^2}{2K} \tag{17}$$

Conclusion

For a convex function f that is L_P-smooth with respect to the P-norm, the preconditioned gradient descent method with step size $\alpha = 1/L_P$ converges in function value to the minimum f^* with a rate of $O(1/K)$. That is, $f(x_K) \to f^*$ as $K \to \infty$.

The proof is completely analogous to the standard proof for GD, but all operations (inner product, norm, gradient, Lipschitz constant) are replaced by their P-analogues.

Base Statements (for second part)

Lemma If f is L–smooth and $\gamma > 0$, then for all $x, y \in \mathbb{R}^d$,

$$f(x - \gamma\nabla f(x)) - f(x) \le -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla f(x)\|^2. \tag{10}$$

If moreover $\inf f > -\infty$, then for all $x \in \mathbb{R}^d$,

$$\frac{1}{2L}\|\nabla f(x)\|^2 \le f(x) - \inf f.$$

General Proof of Convergence of Gradient Descent 2

Theorem Consider the Problem (Differentiable Function) and assume that f is convex and L-smooth, for some $L > 0$. Let $(x_t)_{t\in\mathbb{N}}$ be the sequence of iterates generated by the (GD) algorithm, with a stepsize satisfying $0 < \gamma \le \frac{1}{L}$. Then, for all $x^* \in \arg\min f$, for all $t \in \mathbb{N}$, we have:

$$f(x_t) - \inf f \le \frac{\|x_0 - x^*\|^2}{2\gamma t}.$$

Proof Let f be convex and L–smooth. It follows that

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &= \left\| x_t - x^* - \frac{1}{L}\nabla f(x_t) \right\|^2 \\ &= \|x_t - x^*\|^2 - 2 \cdot \frac{1}{L}\langle x_t - x^*, \nabla f(x_t)\rangle + \frac{1}{L^2}\|\nabla f(x_t)\|^2 \\ &\overset{(1)}{\le} \|x_t - x^*\|^2 - \frac{1}{L^2}\|\nabla f(x_t)\|^2. \end{aligned} \tag{18}$$

Thus, $\|x_t - x^*\|^2$ is a decreasing sequence in t, and consequently

$$\|x_t - x^*\| \leq \|x_0 - x^*\|. \tag{19}$$

Calling upon (10) and subtracting $f(x^*)$ from both sides gives

$$f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \frac{1}{2L}\|\nabla f(x_t)\|^2. \tag{20}$$

Applying convexity we have that

$$\begin{aligned} f(x_t) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^* \rangle \\ &\leq \|\nabla f(x_t)\| \cdot \|x_t - x^*\| \\ &\overset{(19)}{\leq} \|\nabla f(x_t)\| \cdot \|x_0 - x^*\|. \end{aligned} \tag{21}$$

Suppose now that $x_0 \neq x^*$, otherwise the proof is finished. Isolating $\|\nabla f(x_t)\|$ in the above and inserting in (20) gives

$$f(x_{t+1}) - f(x^*) \overset{(20)+(21)}{\leq} f(x_t) - f(x^*) - \frac{1}{2L}\frac{1}{\|x_0 - x^*\|^2}(f(x_t) - f(x^*))^2 \tag{22}$$

Let $\beta = \frac{1}{2L}\frac{1}{\|x_0 - x^*\|^2}$ and $\delta_t = f(x_t) - f(x^*)$. Since $\delta_{t+1} \leq \delta_t$, and by manipulating (22) we have that

$$\delta_{t+1} \leq \delta_t - \beta\delta_t^2 \times \frac{1}{\delta_t\delta_{t+1}} \beta\frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}\delta_{t+1} \leq \delta_t\beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

Summing up both sides over $t = 0, \ldots, T-1$ and using telescopic cancellation we have that

$$T\beta \leq \frac{1}{\delta_T} - \frac{1}{\delta_0} \leq \frac{1}{\delta_T}.$$

Re-arranging the above we have that

$$f(x^T) - f(x^*) = \delta_T \leq \frac{1}{\beta T} = \frac{2L\|x^0 - x^*\|^2}{T}.$$

Proof of Convergence of Gradient Descent with weighted inner product 2

From now on, we will use the following notions:

$$\begin{aligned} \nabla f(x) &= P\nabla_P f(x), \\ P^{-1}\nabla f(x) &= \nabla_P f(x), \\ x_{t+1} &= x_t - \eta\nabla_P f(x_t) \quad \Leftrightarrow \quad x_{t+1} = x_t - \eta P^{-1}\nabla f(x_t). \end{aligned}$$

If you see an inner product written without the subscript P, this is done deliberately and refers to the standard Euclidean inner product.

Proof. Consider the norm induced by P: $\|x\|_P^2 = x^\top Px$. Then the gradient step becomes

$$x_{t+1} = x_t - \eta P^{-1}\nabla f(x_t),$$

which can be written as

$$x_{t+1} - x^* = x_t - x^* - \eta P^{-1}\nabla f(x_t).$$

Taking the squared P-norm of both sides:

$$\begin{aligned}
\|x_{t+1} - x^*\|_P^2 &= \left\|x_t - x^* - \eta P^{-1}\nabla f(x_t)\right\|_P^2 \\
&= \|x_t - x^*\|_P^2 - 2\eta\langle P^{-1}\nabla f(x_t), x_t - x^*\rangle_P + \eta^2\|P^{-1}\nabla f(x_t)\|_P^2 \\
&= \|x_t - x^*\|_P^2 - 2\eta\langle \nabla f(x_t), x_t - x^*\rangle + \eta^2\nabla f(x_t)^\top P^{-1}\nabla f(x_t),
\end{aligned}$$

where we used the identity $\langle u, v\rangle_P = u^\top Pv$ and the fact that $PP^{-1} = I$.

Now, suppose f is convex and L_P-smooth with respect to the P-norm. Then, from standard smoothness inequality:

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L_P}{2}\|x_{t+1} - x_t\|_P^2. \qquad (?)$$

Substitute $x_{t+1} - x_t = -\eta P^{-1}\nabla f(x_t)$:

$$\begin{aligned}
f(x_{t+1}) &\le f(x_t) - \eta\nabla f(x_t)^\top P^{-1}\nabla f(x_t) + \frac{L_P\eta^2}{2}\nabla f(x_t)^\top P^{-1}\nabla f(x_t) \\
&= f(x_t) - \left(\eta - \frac{L_P\eta^2}{2}\right)\nabla f(x_t)^\top P^{-1}\nabla f(x_t).
\end{aligned}$$

Choosing $\eta = \frac{1}{L_P}$, we obtain:

$$f(x_{t+1}) \le f(x_t) - \frac{1}{2L_P}\nabla f(x_t)^\top P^{-1}\nabla f(x_t).$$

From convexity, we also have:

$$f(x_t) - f(x^*) \le \langle \nabla f(x_t), x_t - x^*\rangle.$$

Using Cauchy-Schwarz in P-norm (It was assumed that Cauchy-Schwarz inequality hold in any weighted inner product space? https://math.stackexchange.com/questions/463073/why-does-the-cauchy-schwarz-inequality-hold-in-any-inner-product-spaceLink):

$$\langle \nabla f(x_t), x_t - x^*\rangle \le \|x_t - x^*\|_P \cdot \|P^{-1}\nabla f(x_t)\|_P.$$

Note that:

$$\|P^{-1}\nabla f(x_t)\|_P^2 = \nabla f(x_t)^\top P^{-1}\nabla f(x_t).$$

So we get:

$$f(x_t)-f(x^*) \leq \|x_t-x^*\|_P \cdot \sqrt{\nabla f(x_t)^\top P^{-1} \nabla f(x_t)} \leq \|x_0-x^*\|_P \cdot \sqrt{\nabla f(x_t)^\top P^{-1} \nabla f(x_t)}.$$

Solving for $\nabla f(x_t)^\top P^{-1} \nabla f(x_t)$ and plugging into the earlier bound:

$$f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \frac{1}{2L_P} \cdot \frac{(f(x_t) - f(x^*))^2}{\|x_0 - x^*\|_P^2}.$$

Letting $\delta_t = f(x_t) - f(x^*)$, and $\beta = \frac{1}{2L_P\|x_0-x^*\|_P^2}$, we obtain:

$$\delta_{t+1} \leq \delta_t - \beta \delta_t^2.$$

As in standard analysis, we get ($t = 0, \ldots, T-1$):

$$\delta_{t+1} \leq \delta_t - \beta\delta_t^2 \times \frac{1}{\delta_t \delta_{t+1}} \beta \frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \delta_{t+1} \leq \delta_t \beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

$$\beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \quad \Rightarrow \quad T\beta \leq \frac{1}{\delta_T} - \frac{1}{\delta_0} \leq \frac{1}{\delta_T},$$

which implies:

$$f(x_T) - f(x^*) = \delta_T \leq \frac{1}{\beta T} = \frac{2L_P\|x^0 - x^*\|_P^2}{T}.$$